DOCUMENT RESUME

ED 455 293 TM 033 147

AUTHOR Abedi, Jamal
TITLE Validity Considerations in the Assessment of LEP Students Using Standardized Achievement Tests.
PUB DATE 2001-04-00
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Correlation; Elementary Secondary Education; *English (Second Language); Language Proficiency; *Limited English Speaking; *Standardized Tests; *Validity
IDENTIFIERS Iowa Tests of Basic Skills; Language Assessment Battery; Latent Variables; Stanford Achievement Tests

ABSTRACT

Validity and reliability issues in standardized testing of students of limited English proficiency (LEP) were studied. Existing data from four different school sites were obtained for LEP and non-LEP students for three different standardized tests, the Stanford Achievement Tests (Ninth edition), the Iowa Tests of Basic Skills, and the Language Assessment Scale. Several different analyses were performed on the available data, including descriptive statistics by LEP status, analyses of internal consistency of the test items by LEP status, and analyses comparing the structural relationships of the instruments across LEP categories. Analysis results are consistent with the literature and indicate that: (1) student English language proficiency is associated with performance on content-based assessments; (2) there is a performance gap in content assessment between LEP students and their native English-speaking peers; and (3) the performance gap between LEP students and non-LEP students increases as the language load of the assessment tools increased. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 455 293

Validity considerations in the assessment of LEP students using standardized achievement tests[1]

Jamal Abedi

University of California, Los Angles

National Center for Research on Evaluation, Standards, and Student Testing

(CRESST)

Paper presented at the 2001 Annual Meeting of the American Educational Research Association, Seattle, Washington

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
J. Abedi
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

TM033147

[1] For a detailed description of these studies see Abedi and Leon (1999); Abedi, Leon, and Mirocha (2000).

BEST COPY AVAILABLE

Perspective

Recent legislation calls for inclusion of all students including those with disability and limited English language proficient (LEP) students. Innovative ways of assessing student performance are encouraged, including modifications to existing instruments for English language learners (August & Hakuta, 1997). This call has prompted new interest in modifying assessments to "accommodate" Students with disabilities and English language learners, to enhance the validity and equitability of the inferences drawn from the assessments themselves.

However, as most standardized, content-based tests are conducted in English and normed on native English speaking test populations, they may function as English language proficiency tests. Students with limited English proficiency (LEP) may be unfamiliar with scriptally implicit questions, may not recognize vocabulary terms, or may mistakenly interpret an item literally (Duran, 1989; Garcia, 1991). Results of analyses on data sets from several large school districts nationwide have raised issues on the use of such tests for ELL students. For example, the results of analyses on the standardized achievement tests have indicated that language factors act as sources of measurement error on the test and may be a source of construct irrelevant variance. In this presentation issues concerning validity, reliability and linguistic factors will be discussed.

Procedure

Existing data from four different school sites nationwide were obtained. To assure anonymity, these data sites will be referred to as Sites 1 to 4. Site 1 is a large urban school district that provided ITBS performance data from 1999 for grades 3 through 8. In addition to ITBS data, student background data were also provided. Site 2 is a state with a very large number of English language learners. We gained access to the Stanford 9 test data for all students in Grades 2 to 11 who were enrolled in the statewide public schools for the 1997-1998 academic year. These data included student responses to test items (item-level data), subsection scores, and student background data. The background data included gender, ethnicity, free/reduced-price lunch participation, parent education,

student LEP status, and Students with Disabilities (SD) status. Site 3 is an urban school district. Stanford 9 test data were available for all students in Grades 10 and 11 for the 1997-1998 academic year. These data included student responses to test items (item-level data), subsection scores, student background data, and test accommodation data. Site 4 is a state with a significant number of English language learners. The Department of Education in this state gave us access to the Stanford 9 summary test data for all students in Grades 3, 6, 8, and 10 who were enrolled in the statewide public schools for the 1997-1998 academic year. Item-level data was available for a sample of this states population for grades 3, 5, 7, and 9 from the 1998-1999 academic year. Student background data were also available form this site and included gender, ethnicity, free/reduced-price lunch participation, student LEP status, and Students with Disabilities (SD) status.

There were similarities and differences among the four data sites. The sites were similar in that they all used standardized tests for measuring student school achievement in English and other content-based areas, they all had an index of students' English language proficiency status (LEP or bilingual status), and they all contained some student background information. However, they differ in the type of standardized achievement tests, the index of English language proficiency status, and the type of background variables that they provided for their students. These differences may limit our ability to perform identical analyses at the different sites for cross validation purposes. However, there were enough similarities in the data structures at the four different sites to allow for interesting and valid comparisons. For example, data on students' LEP status were provided by three of the four sites; one site provided information on student "bilingual status" rather than LEP status.

The standardized tests that were used in the four sites were the Stanford Achievement Test Series, Ninth Edition (Stanford 9), the Iowa Tests of Basic Skills (ITBS), and the Language Assessment Scale (LAS). Among the background data that were provided by the sites are race, gender, birth date, and number of years of participation in a bilingual education program (number of years of bilingual service).

Descriptive statistics comparing LEP and non-LEP student (or LEP and non-

LEP or bilingual and non-bilingual) performance by subgroup and across the different content areas revealed major differences. Included in the descriptive statistics section was a Disparity Index (the Disparity Index of non-LEP over LEP students). This index showed major differences between students with different language backgrounds. However, the more English language complexity involved in the assessment tool, the greater was the disparity index.

In multiple regression models, student LEP status was related to student test scores and background variables. In a canonical correlation model the relationship between student English language proficiency level, parent education, and family SES (the SET 2 variables) and Stanford 9 performance (the Set 1 variables) was examined. The results of these analyses confirmed our earlier findings that the higher the English "language load" in the assessment, the larger the gap between performance of LEP and non-LEP students.

The term "language load" refers to the linguistic complexity of the test items. In her language analysis of standardized achievement tests, Bailey (2000) uses the term "language demand" and indicated that the language demand of standardized achievement tests could be a potential threat to the validity of these tests when administered to English language learners. Because of this source of threat, she added, the assessment many not present an accurate picture of LEP student content knowledge. Bailey elaborated on the concept of language demand as uncommon vocabulary, non-literal usage (idioms), complex or atypical syntactic structure, uncommon genre, or multi-clausal processing. For this study, we did not perform any linguistic analyses of test items. However, test items in some content areas involve more English language demand than in other content areas. For example, it is obvious that in reading assessments there is more English language load involved than in other content-based areas such as math and science.

Several different analyses were performed on the available data, including descriptive statistics by LEP status, analyses of internal consistency of the test items by LEP status, and analyses comparing the structural relationships of the instruments across various LEP categories. Descriptive analyses show that LEP students generally perform at a lower level than non-LEP students on reading,

science, and math subtests—a strong indication of the impact of English language proficiency on assessment. However, the level of impact of language proficiency on assessment of LEP students is greater in the content areas with high language demand. For example, analyses show that LEP and non-LEP students have the greatest performance differences in reading. The gap between the performance of LEP and non/LEP students becomes smaller in other content areas with less language demand. The difference between LEP and non-LEP students' performance becomes smallest in math, where language has less of an impact on the assessment.

Table 1 summarizes descriptive statistics for Site 3. As data in Table 1 indicated, the performance-gap between LEP and non-LEP decreases as we move from reading to science and from science to math. That is, the performance gap between LEP and non-LEP is substantially higher in subject areas with higher language load. For example, the mean NCE score for reading in grade 10 for LEP is 24.0 as compared with a mean of 38.0 for non-LEP, a difference of 14 NCE points. The mean science score is 32.9 for LEP and 42.6 for non-LEP with a difference of about 10 points, substantially less than the difference in reading. For math, mean for LEP is 36.8 and for non-LEP the mean is 39.6, a difference of about 2 points.

Note. LEP = limited English proficient. SD = students with disabilities.

Table 1 also presents means and standard deviations for students in Grade 11. The mean score for reading for all students in Grade 11 is 36.2 with a standard deviation of 19.0. For science, the mean score for all students is 38.2 (*SD* = 18.9), and for math, the mean score is 44.0 (*SD* = 21.2). These results are very similar to those obtained for students in Grade 10. As discussed in the previous section, the means of subscale scores increase as we move from reading to science and from science to math. For science, there was a 6 score-point increase over reading (.4 standard deviation), and for math, there was a 23 score-point increase (1.5 standard deviation) over reading and a 17 score-point increase (1.1 standard deviation) over science. This trend of increase in subscale score is due to several factors including content and language factors. The language factors are particularly important for the LEP group.

To present a more clear picture of differences between the performance of LEP and non-LEP students, Disparity Index (DI) was computed for data in site 3. Table 2 presents the DI s. The Disparity Indices (an index of performance differences between LEP and non-LEP) shown in Table 1.21 suggest that the higher the level of language load in the assessment, the larger is the gap between the performance of LEP and non-LEP. For example, for both grade 10 and 11 students, the DI is largest for reading (58.3 for grade 10 and 70.7 for grade 11), becomes smaller for science (29.5 for grade 10 and 39.4 for grade 11) and becomes almost zero for math (7.6 for grade 10 and –0.7 for grade 11).

The results of our analyses also indicate that test items for LEP students, particularly LEP students at the lower end of the English proficiency spectrum, suffer from lower internal consistency. That is, the language background of students may add another dimension to the assessment, a language dimension. Thus, we speculated that language might act as a source of measurement error in such cases.

These findings are consistent across the grade level and across the different data sites. Table 3 is an example of this comparison. As data in Table 3 show, alpha coefficients are generally lower for LEP students. For example, for Vocabulary subscale, alpha for non-LEP high SES is .828 as compare with alpha of .666 for LEP students. Similar trend can be seen in all other subject areas in Table 3.

We also compared LEP and non-LEP students on individual test items. We categories test items based on the index of difficulty (proportion of correct responses) into three categories, *small, moderate* and *large* differences. A small difference was considered as less than 9 percentage points. A moderate difference was considered as 10 to 20 percentage points. A large difference was considered to be greater than 20 percentage points. Differences between LEP and non-LEP students were substantially higher in subject areas with higher language load. We reported the differences between LEP and non-LEP in three categories, for all LEP, for LEP non-accommodated and for LEP accommodated. Our previous analyses indicated that LEP accommodated had the lowest level of language proficiency, thus, they were accommodated. For example, in grade 10, the proportion of difference for LEP accommodated is 59% in difficult items in

reading. This difference decreased to 22% in science and completely disappears in math. Once again, these analyses point to the impact of language on test items.

Analyses of the structural relationships between individual items and between items with the total test scores showed a major difference between LEP and non-LEP students. Structural models for LEP students demonstrated lower statistical fit. Further, the factor loadings were generally lower for LEP students and the correlations between the latent content-based variables were weaker for LEP students.

To compare within-test and cross-test structural relationships between LEP and non-LEP students, a series of simple structure confirmatory models were created. In creating these models, test items in each of the three content areas (reading, science, and math) were grouped as "parcels." Correlation between the reading, math and science latent variables were estimated. Models were tested on randomly selected sample populations to demonstrate the consistency of the results.

As the results show, correlations of item parcels to the latent factors are consistently lower for LEP students than they are for non-LEP students. This finding was true for all parcels regardless of which grade or which sample of the population was tested. For example, in grade 9 for LEP students the correlation for the four reading parcels ranged from a low of .719 to a high of .779 across the two samples as shown in table 4.11. In comparison, for non-LEP students the correlation for the four reading parcels ranged from a low of .832 to a high of .858 across the two samples. The item parcel correlations were also larger for non-LEP students then for LEP students in math and science. Again these results were consistent across the different samples. The paired correlations between the latent factors were also larger for non-LEP students then they were for LEP students. This gap in latent factor correlations between non-LEP and LEP students was especially large when there was a larger language demand difference on the test items. For example, in the grade 9 sample population #1 the correlation between latent factors for math and reading for non-LEP students was .782 compared to a correlation of .645 for LEP students. When comparing the

latent factor correlations between reading and science from the same population the correlation was still larger for non-LEP students (.837) than for LEP students (.806), but the gap between the correlations decreased. This is likely due to a larger language demand difference between the reading and math tests as compared to the reading and science tests.

Multiple group structural models were run to test whether the differences between non-LEP and LEP students mentioned above were significant. There was significant differences for all constraints tested at the $p<.05$ level. These findings are consistent with the literature, which suggests that English language proficiency may impact assessment for LEP students.

The results of our analyses of data from the four sites were consistent with the literature and indicated that:

a. Student English language proficiency level is associated with performance on content-based assessments.
b. There is a performance gap in content assessment between LEP students and their native English-speaking peers (non-LEP students).
c. The performance gap between LEP students and non-LEP students increases as the language load of the assessment tools increases.

References

Abedi, I. Leon, S. & Mirocha, J. (2000). *Examining ELL and non-ELL student performance differences and their relationship to background factors: Continued Analyses of Extant Data*. University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, I. Leon, S. (1999). *Impact of students' language background on content-based performance: Analyses of extant data*. University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

August, D. & Hakuta, K. (Eds.). (1997). *Improving schooling for language-minority chi'dren: A research agenda*. Washington, D.C.: National Academy Press.

Durán, R. P. (October 1989). Assessment and instruction of at-risk Hispanic students. *Exceptional Children, 56*(2), 154-158.

Bailey, A. (2000). *OBEMLA Report*. University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.Garcia, G.E. (1991). Factors influencing the English Reading Test Performance of Spanish-speaking Hispanic Children. *Reading Research Quarterly, 26*(4), 371-391.

Table 1. Normal Curve Equivalent Means and Standard Deviations for Students in Grades 10 and 11, Site 3 School District

	Reading		Science		Math	
	M	*SD*	*M*	*SD*	*M*	*SD*
Grade 10						
SD only	16.4	12.7	25.5	13.3	22.5	11.7
LEP only	24.0	16.4	32.9	15.3	36.8	16.0
LEP & SD	16.3	11.2	24.8	9.3	23.6	9.8
Non-LEP & SD	38.0	16.0	42.6	17.2	39.6	16.9
All students	36.0	16.9	41.3	17.5	38.5	17.0
Grade 11						
SD Only	14.9	13.2	21.5	12.3	24.3	13.2
LEP Only	22.5	16.1	28.4	14.4	45.5	18.2
LEP & SD	15.5	12.7	26.1	20.1	25.1	13.0
Non-LEP & SD	38.4	18.3	39.6	18.8	45.2	21.1
All Students	36.2	19.0	38.2	18.9	44.0	21.2

Table 2. SITE 3 Disparity Index (DI) Non-LEP/Non SD Students compared to LEP Only

Grade	Disparity Index (DI)			
	Reading	Math Total	Math Calculation	Math Analytical
3	53.4	25.8	12.9	32.8
6	81.6	37.6	22.2	46.1
8	125.2	36.9	25.2	44.0

Table 3. Site 2 Stanford 9 Sub-scale Reliabilities (1998) GRADE 9 Unadjusted Alpha's

Sub-scale(Items)	Non-LEP Students Hi SES	Low SES	English Only	FEP	RFEP	LEP
Reading	N=205,092	N=35,855	N=181,202	N=37,876	N=21,869	N=52,720
-Vocabulary (30)	.828	.781	.835	.814	.759	.666
-Reading Comp. (54)	.912	.892	.916	.903	.877	.833
Average reliability	.870	.837	.876	.859	.818	.750
Math	N=207,155	N=36,588	N=183,262	N=38,329	N=22,152	N=54,815
-Total (48)	.899	.853	.898	.898	.876	.802
Language	N=204,571	N=35,886	N=180,743	N=37,862	N=21,852	N=52,863
-Mechanics (24)	.801	.759	.803	.802	.755	.686
-Expression (24)	.818	.779	.823	.804	.757	.680
Average reliability	.810	.769	.813	.803	.756	.683
Science	N=163,960	N=28,377	N=144,821	N=29,946	N=17,570	N=40,255
-Total (40)	.800	.723	.805	.778	.716	.597
Social Science	N=204,965	N=36,132	N=181,078	N=38,052	N=21,967	N=53,925
-Total (40)	.803	.702	.805	.784	.722	.530

Table 3. Site 3 School District Item Level Data: Raw Score *P*-Value Difference with Non-LEP students as a Reference—Reading, Science, and Math Stanford 9 Scores, Grades 10 and 11

	Percent of Items with Small, Moderate & Large *p*-value differences*								
	Reading (54 Items)			Science (40 Items)			Math (48 Items)		
	Small	Mod.	Large	Small	Mod.	Large	Small	Mod.	Large
Grade 10									
All LEP	18%	54%	28%	88%	10%	2%	100%	0%	0%
Non-Accom.	54%	44%	2%	95%	5%	0%	100%	0%	0%
Accom.	11%	30%	59%	68%	22%	10%	88%	12%	0%
Grade 11									
All LEP	11%	56%	33%	73%	23%	5%	98%	2%	0%
Non-Accom.	37%	52%	11%	85%	10%	5%	100%	0%	0%
Accom.	4%	30%	67%	68%	20%	13%	90%	10%	0%